Building API-Based Web Archiving Systems and Services

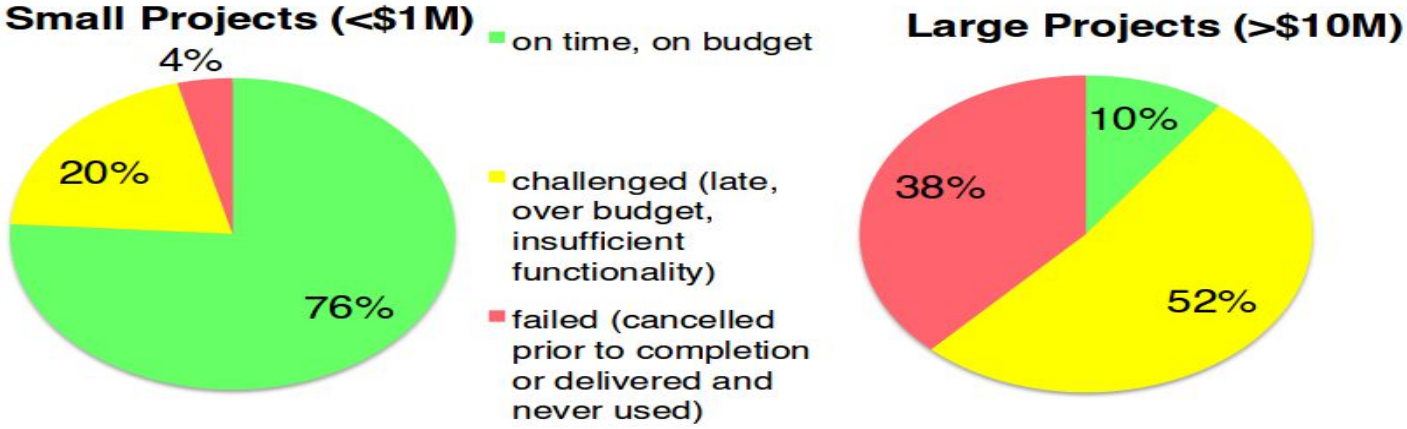
David Rosenthal, LOCKSS | Stanford (http://blog.dshr.org/)
Jefferson Bailey, Internet Archive (@jefferson_bail)
Nicholas Taylor, Stanford University (@nullhandle)

Why do we need APIs?

David S. H. Rosenthal

LOCKSS Program
Stanford University Libraries
http://www.lockss.org/
http://blog.dshr.org/

Small vs. Large Software-Development Projects



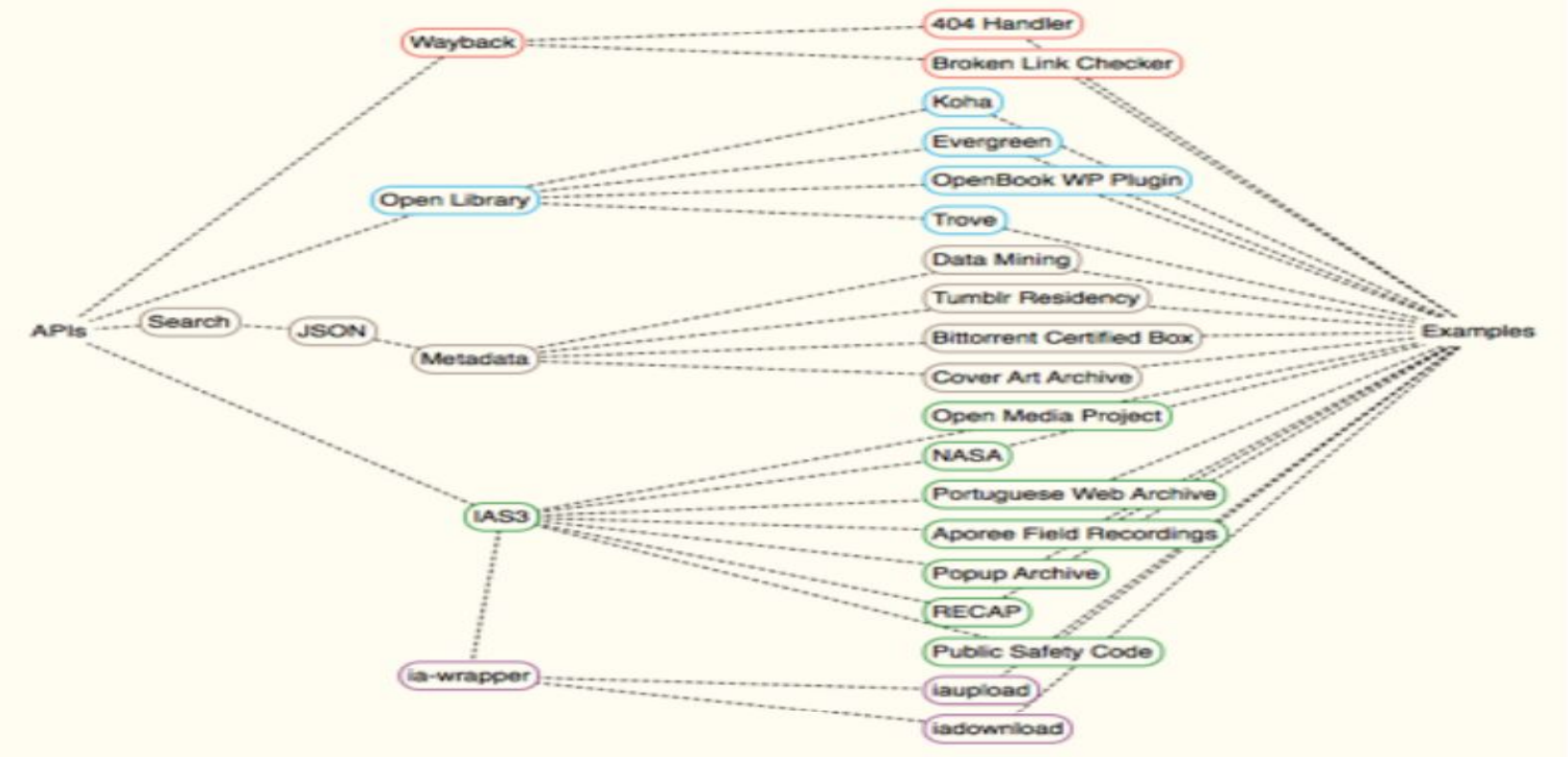
Do Big Project via Service-Oriented Architecture + Many Small Services!

CHAOS MANIFESTO 2013 Think Big, Act Small, www.standishgroup.com. Based on the collection of project case information on 50,000 real-life IT environments and SW projects. Surveying since 1985.

CEO: Amazon shall use SOA! (2002, 7 years after started company)

1. "All teams will henceforth expose their data and functionality through service interfaces."
2. "Teams must communicate with each other through these interfaces."
3. "There will be no other form of interprocess communication allowed: no direct linking, no direct reads of another team's data store, no shared-memory model, no back-doors whatsoever."
4. "Service interfaces must be designed from the ground up to be externalizable. That is to say, the team must plan and design to be able to expose the interface to developers in the outside world. "
5. "Anyone who doesn't do this will be fired."

Internet APIs + Uses



ARCHIVE-IT

STANFORD UNIVERSITY LIBRARIES

LOCKSS

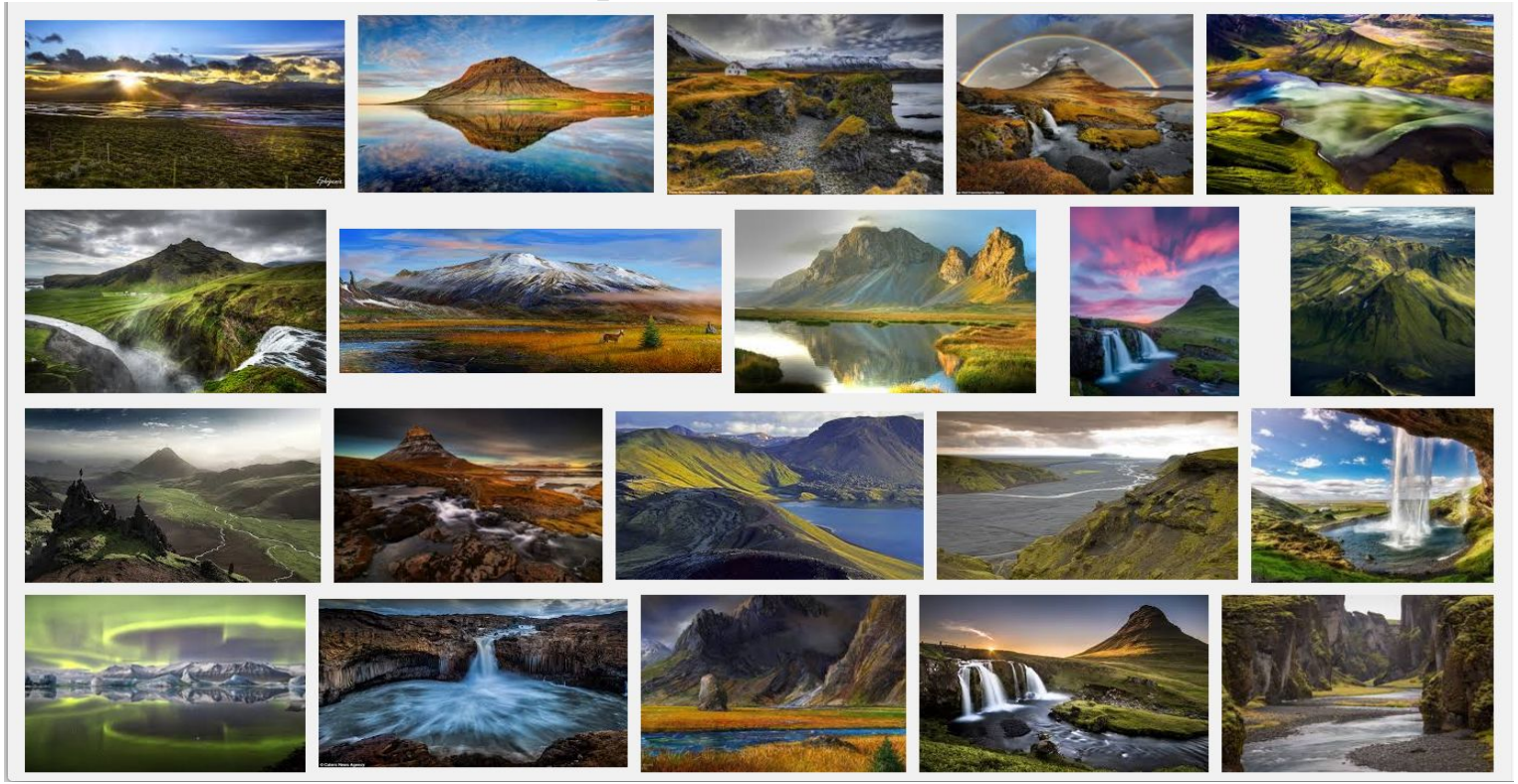
WASAPI

Credit: Alexis Rossi

The Big Picture

- Ingest:
  - Sharing Crawlers, Capturing renderings, Deduplication
  - Divide/Conquer crawling, Soft Errors, Metadata extraction
  - Crawl management
- Preservation:
  - Detect/repair damage, advertise holdings
- Dissemination:
  - Memento, federated browsing, text & metadata search
  - Bulk access, format migration, data mining
  - Emulation

Landscapes & WASAPI



Jefferson Bailey, Internet Archive (@jefferson_bail)

Growth in Web Archiving (NDSA & Archive-It)

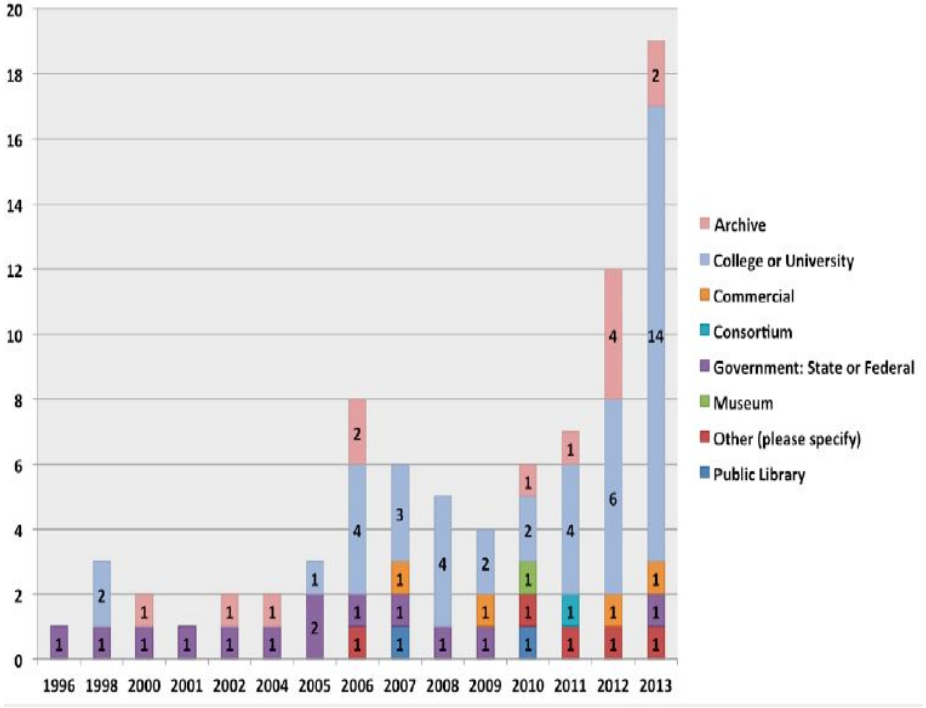
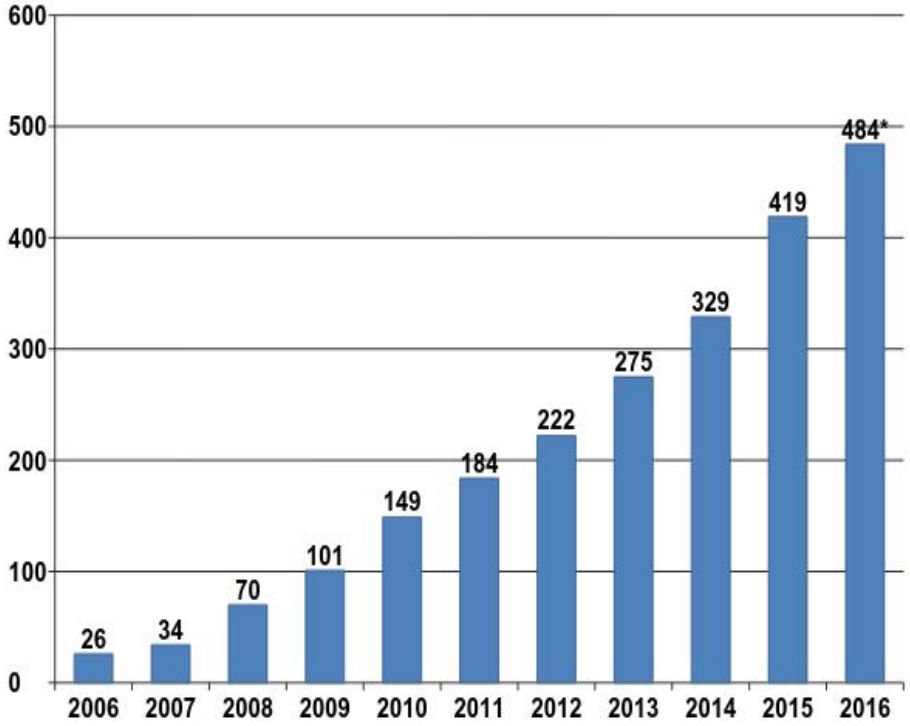


FIGURE 4: YEAR INSTITUTIONS BEGAN ARCHIVING WEB CONTENT



ARCHIVE-IT

STANFORD UNIVERSITY LIBRARIES

LOCKSS

WASAPI

Local Preservation of Web Archives

Recent Surveys of local preservation of web data

- NDSA: 18%-20% (2011, 2013, 2016)
- AIT: 20% of respondents (2016)
- Reasons include
  - No local preservation plan
  - Trust in service
  - Doesn't integrate with existing workflows
  - Too much data

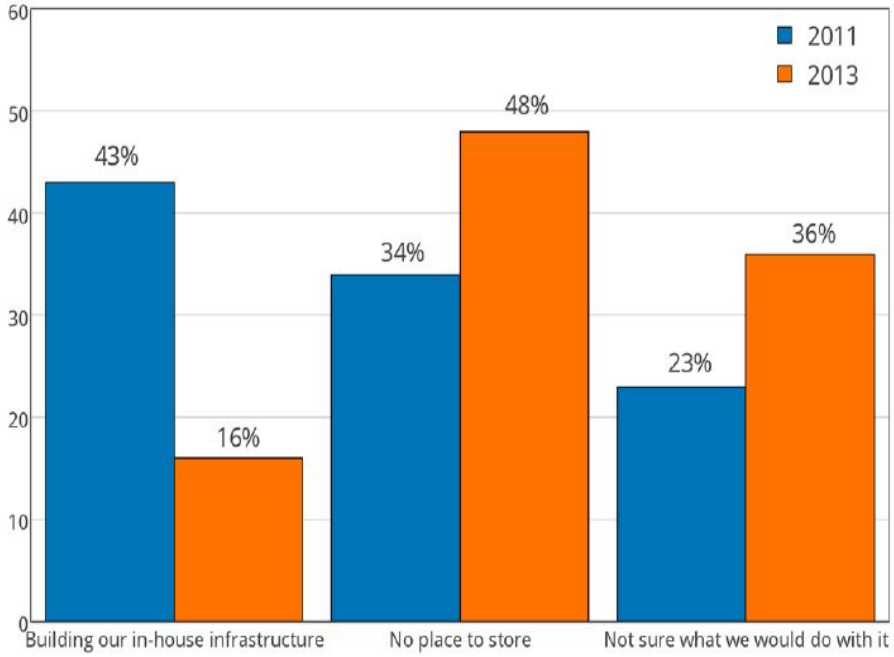


FIGURE 15: REASONS FOR NOT TRANSFERRING DATA FROM AN EXTERNAL SERVICE

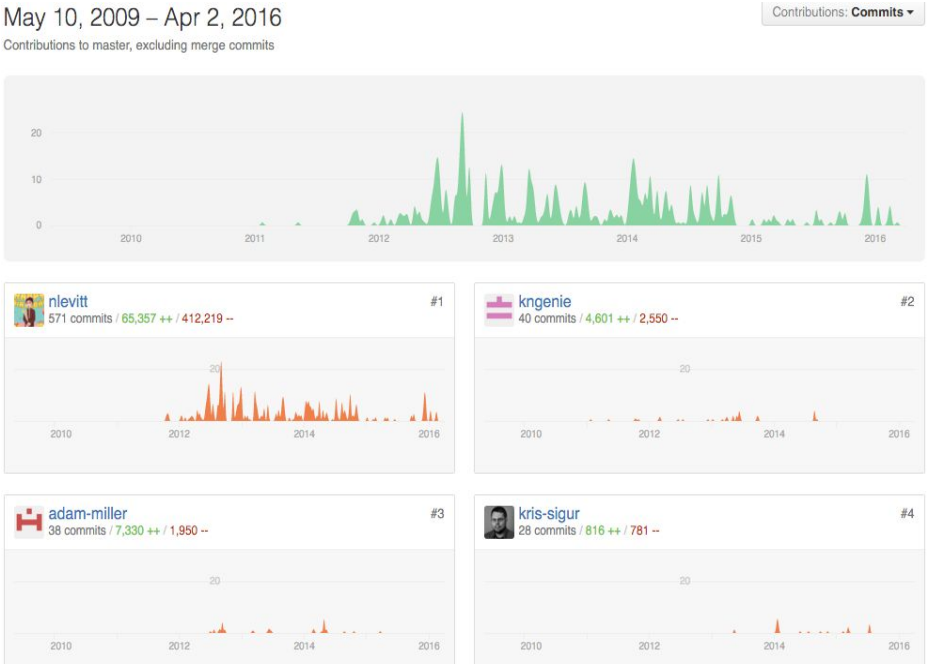
ARCHIVE-IT

STANFORD UNIVERSITY LIBRARIES

LOCKSS

Community Involvement in WA Development

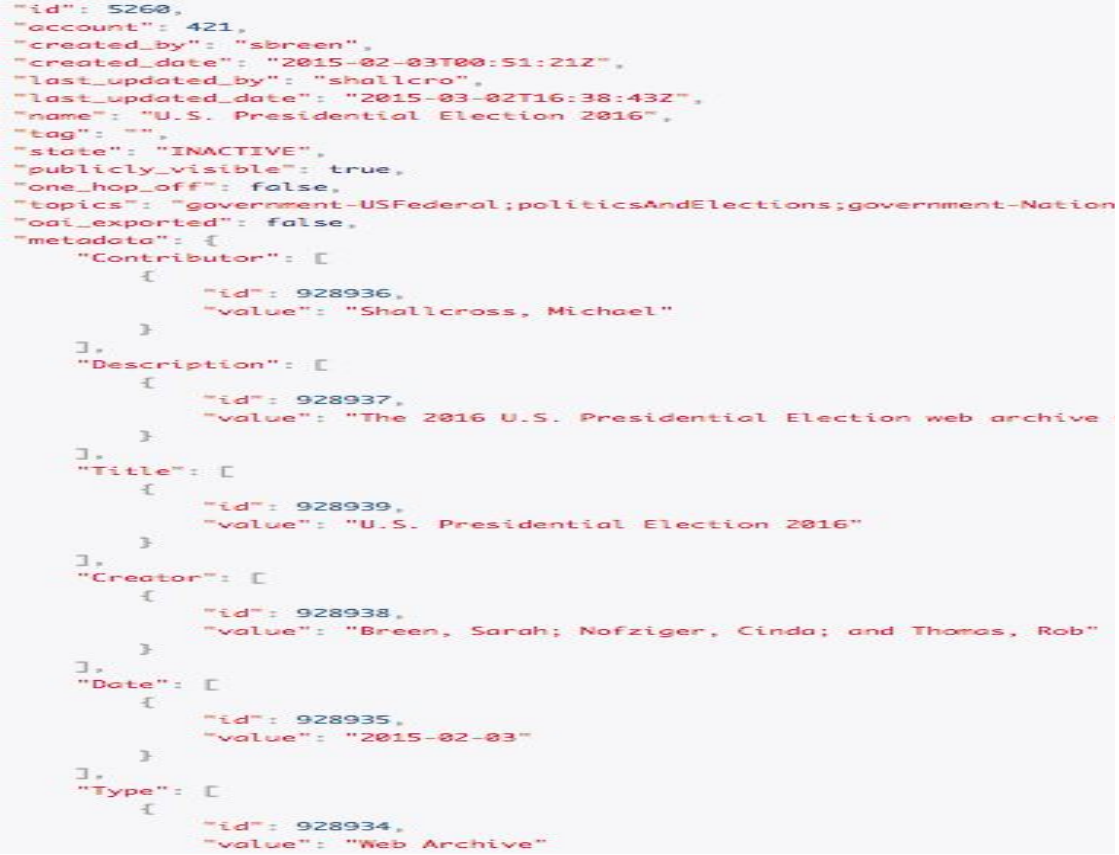
- Few coordinated efforts on shared tools
- Historical reliance on few providers
- Variance of coordination on emergent efforts & foresight on interoperability
- Few on-ramps for not-dev participation
- Yet some collaborative digital library efforts have proven successful
- Emergence of broader web archiving community of practice



Other Challenges

- Web Archiving often still a niche collecting activity
- Use largely TBD or not measured
- Convenience of end-to-end services diminishes tech needs
- Little familiarity with formats, software, or processes
- Nascent community impetus to join or advise on broad technical development activities

ARCHIVE-IT STANFORD UNIVERSITY LIBRARIES LOCKSS



- Wayback APIs
- Archive-It Partner Metadata APIs
- Data Analytics APIs (crawl logs and reports)
- Index (CDX) APIs
- Upload APIs (non-web)
- Internal APIs

https://github.com/ArchiveLabs/api.archive.org

ARCHIVE-IT STANFORD UNIVERSITY LIBRARIES LOCKSS

WASAPI: Web Archiving Systems APIs

- "Systems Interoperability and Collaborative Development for Web Archives"
  - National Leadership Grant, National Digital Platform, R&D
  - IA/AIT (PI), Stanford, UNT, Rutgers
  - 2-year project started January 2016
  - National Symposium Early 2017

WASAPI: Web Archiving Systems APIs

Three Key Areas of R&D:

1) What are the attributes of a community model that can support sustainable and broad-based collaborative web archiving technology development?
2) What are the community needs and downstream uses for the planned Export APIs (by AIT & LOCKSS) to facilitate transfer of web archive data between distributed systems and what other prospective APIs does it point to?
3) How can better interoperability of web archiving systems support new forms of access and research use?

INSTITUTE of Museum and Library SERVICES

STANFORD UNIVERSITY

LOCKSS LOTS OF COPIES KEEP STUFF SAFE

NORTH TEXAS

RUTGERS THE STATE UNIVERSITY OF NEW JERSEY

ARCHIVE-IT

STANFORD UNIVERSITY LIBRARIES

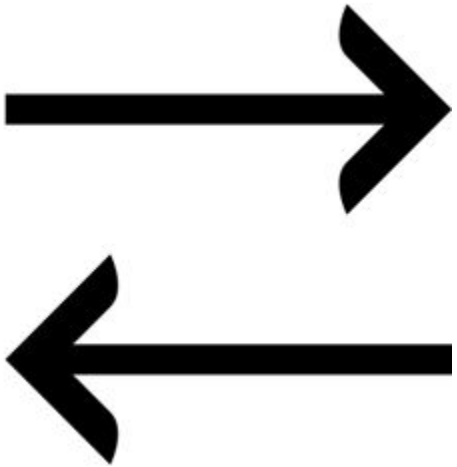
LOCKSS

WASAPI: Web Archiving Systems APIs

Outcomes:

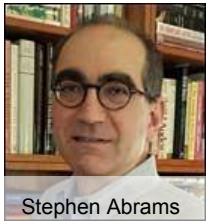
1) Seed & launch a community modeled on the characteristics of successful development and participation communities ID'ed by project
2) Build WARC & derivative dataset APIs (AIT & LOCKSS) and test via transfer to partners (SUL, UNT, Rutgers) to enable better distributed preservation and access
3) Sketch a blueprint and technical model for future web archiving APIs informed by R&D
4) Seed a technical infrastructure that will facilitate more computational and distributed research use of web archive collections

WASAPI Technical Working Group and Current Progress



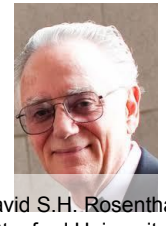
Nicholas Taylor (@nullhandle)
Web Archiving Service Manager
Stanford University Libraries

Technical Working Group



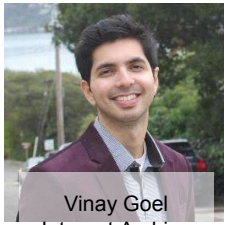
Stephen Abrams
California Digital Library

Andy Jackson
British Library



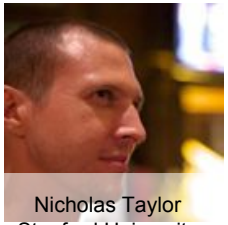
David S.H. Rosenthal
Stanford University

Jefferson Bailey
Internet Archive



Vinay Goel
Internet Archive

Courtney Mumma
Internet Archive



Nicholas Taylor
Stanford University

Tom Cramer
Stanford University

Mark Phillips
University of North Texas

Matt Weber
Rutgers University

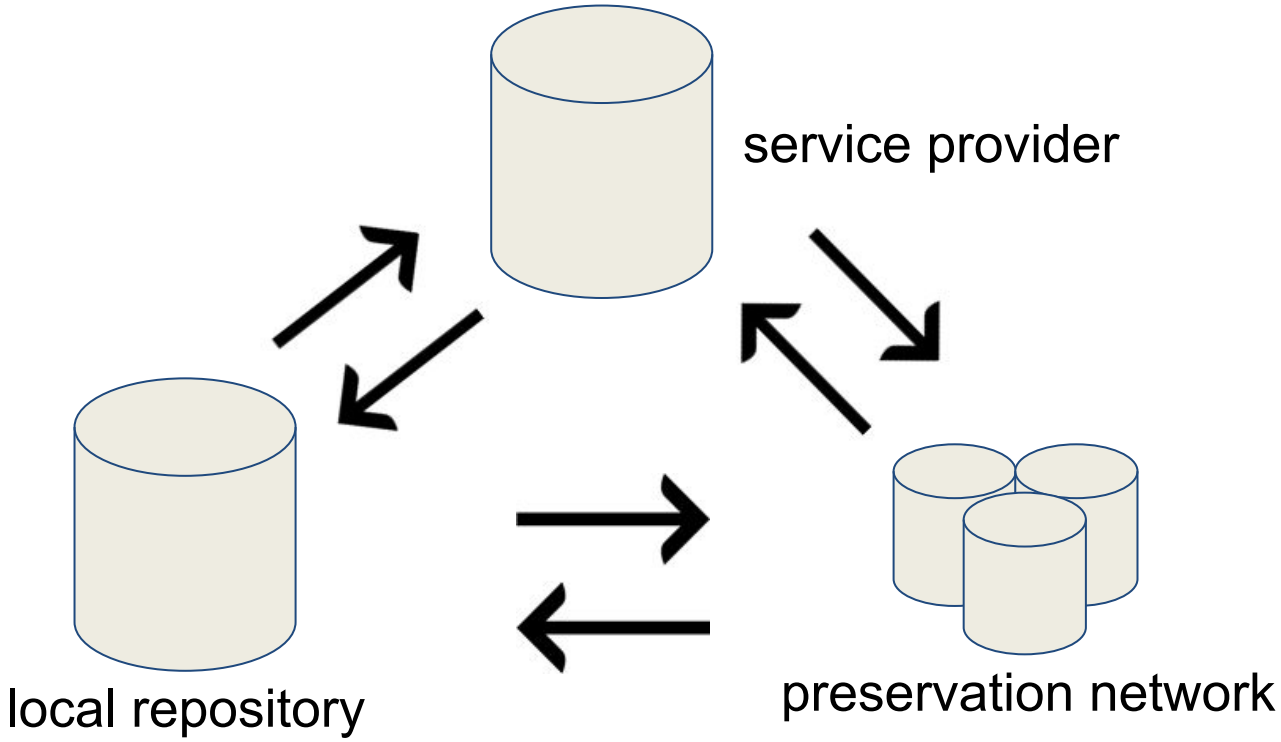
related API work

- CDX Server API (IA, IIPC)
- derivative formats (Archive-It, BL)
- crawl logs/partner data (Archive-It)
- Wayback Machine APIs (IA)
- proliferating capture tools (GWU, IA, Rhizome)
- Cobweb (CDL, Harvard, UCLA)

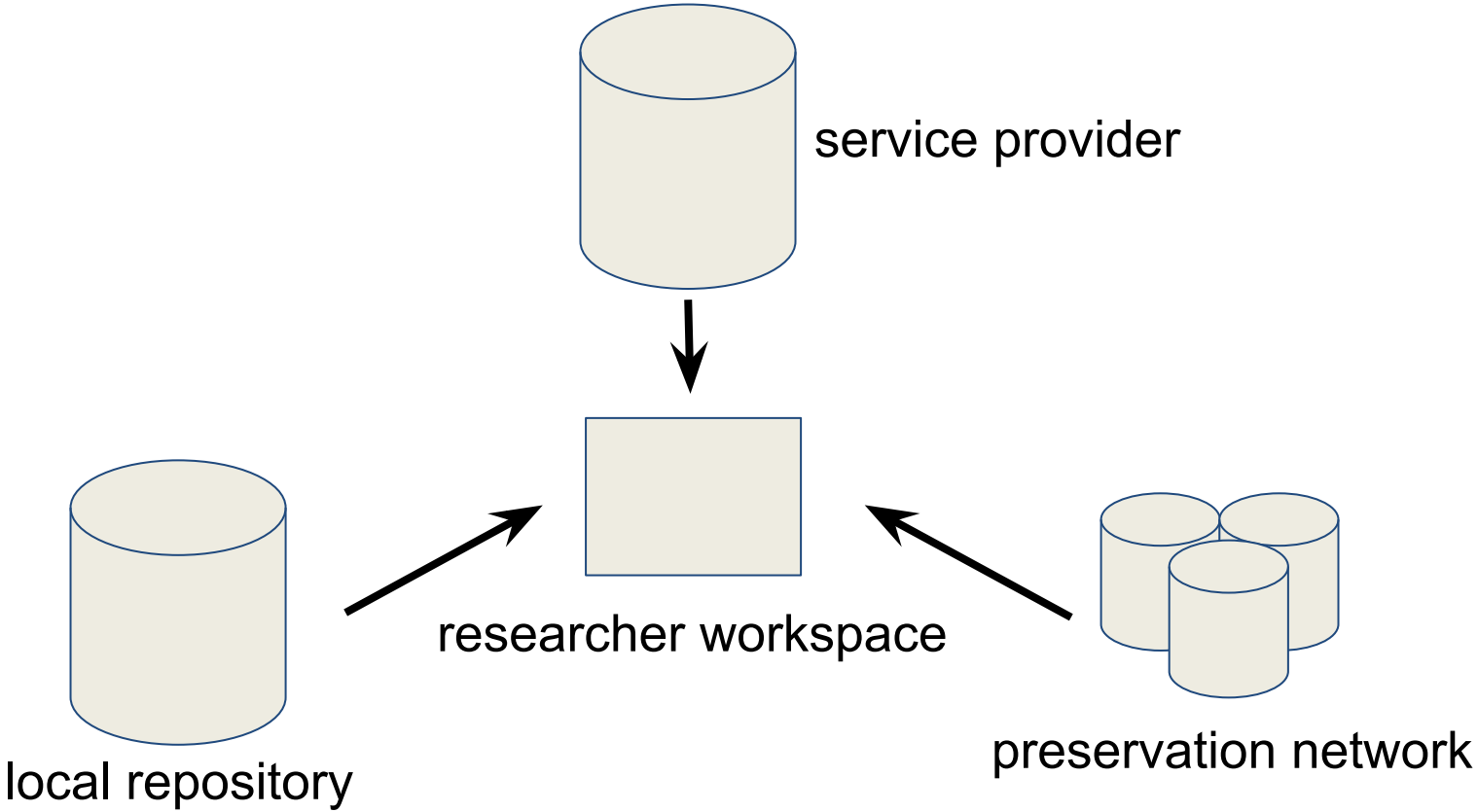
use cases

- Archive-It →
  - **partner IR/local use**
  - DPN
  - LOCKSS (PLN)
- CDL → Archive-It (migration)
- DLSS → IA (WebBase)
- [EoT partners] ← → [EoT partners]
- IA global Wayback→
  - LOCKSS (OA content)
  - national libraries
- LOCKSS (.gov) → IA
- [any web archive] →
  - **researcher**
  - original publisher

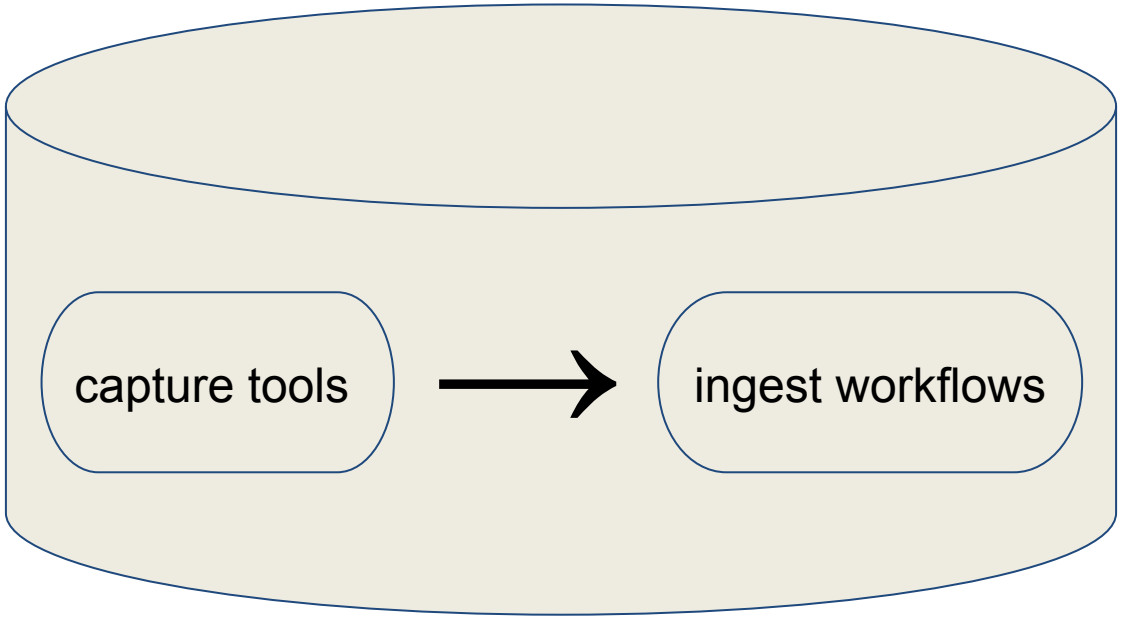
data exchange b/t repositories



standardizing researcher data access



data exchange within repositories



candidate features discussed

- content negotiation for W/ARC or derivatives
- protocol negotiation for transfer handoff
- ability to specify parameters for custom export
- metadata for provenance, crawler configuration, crawl logs, description
- request custom data extraction
- authentication + privileges management

export API example

- authentication
  - (system tracks permissions)
- submit institution ID
  - return associated collection IDs
- submit collection ID(s)
  - return associated job IDs
- submit job ID(s)
  - return associated W/ARC files
- submit candidate W/ARC files
  - return supported protocols
- initiate transfer
  - (transfer files)
  - (acknowledge transfer completion status)

THANKS! (discussion is next)

WASAPI

https://groups.google.com/forum/#!forum/wasapi-community

https://github.com/WASAPI-Community

https://www.imls.gov/sites/default/files/proposal_narritive_lg-71-15-0174_internet_archive.pdf

Nicholas Taylor, Stanford University (@nullhandle)
Jefferson Bailey, Internet Archive (jefferson@archive.org)
David Rosenthal, LOCKSS | Stanford University

Discussion Questions

- What APIs have attendees built, or are currently using, in their web archiving activities?
- Are these APIs RESTful? If not, why not?
- What frameworks/languages were they built with? What are other notable characteristics of their development and maintenance?
- What part of the web archiving lifecycle would most benefit from next-stage API development, post-WASAPI?